**基于集成学习的最小错误率训练算法**

陈昉[[1]](#footnote-2)，王志豪2，赵程绮1，李江梦1

（1 厦门大学 软件学院，福建 厦门 361005

 2 厦门大学 自动化系，福建 厦门 361005）

**摘要**：最小错误率训练是统计机器翻译的标准调参方法，在统计机器翻译建模过程中发挥着重要作用。然而，该方法在训练过程中容易出现训练过拟合现象，即开发集训练得到的权重无法很好地适用于翻译测试集。针对该问题，本文引入集成学习方法来优化调参。具体而言，在调参时本文挑选不同的特征子集来训练多组特征权重，并计算权重之间的空间距离以删除不合理的特征权重，最后根据各组子集在开发集上的BLEU值来进行加权平均，获得最终的特征权重。NIST和IWSLT实验结果表明，本文方法具有较好的效果。

**关键词：**统计机器翻译；最小错误率训练；集成学习

**Minimum Error Rate Training based on Ensemble Learning**

Chen Fang1, WANG Zhi-hao2, Zhao Cheng-qi1, Li JIANG-meng1

(1 *Software School of Xiamen University, Xiamen University, Xiamen, Fujian* 361005*, China;*

2 *Automation Department of Xiamen University, Xiamen University, Xiamen, Fujian* 361005*, China*)

**Abstract:** Minimum error rate training (*MERT*) is a standard tuning parameter procedure in statistical machine translation, playing a significant role in the process. However, overfitting phenomenon is likely to occur in the original *MERT*, that is, the weights trained from development set cannot be fit for test set. In view of this issue, we adopt ensemble learning method to the training process in this paper. To be specific, we first select different feature subsets to acquire several groups of feature weights through *MERT*, and then filter out unreasonable weights according to their spatial distance, and at last we compute the weighted average as the final feature weight based on their *BLEU* scores on development set. Experiments on *NIST* and *IWSLT* show that our method is efficient for the translation tasks using the training and testing data sets of different domains.

**Key words:** Statistical Machine Translation; Minimum Error Rate Training; Ensemble Learning

**1 引言**

 近年来，统计机器翻译[1,2]（*Statistical Machine Translation*）研究取得了巨大的进展，出现了许多翻译模型，并且有些模型已经被用于商业产品的实用化之中，方便了人们的日常生活。现有的统计机器翻译都是以对数线性模型[3]（*Log-linear Model*）为基本框架来进行建模的。随着该模型的引入，翻译模型的特征选择变得更加灵活。然而，如何计算每个特征的权重问题也随之产生。该问题成为了统计机器翻译研究的一个重要问题。

 针对上述问题，现有研究大多采用最小错误率训练[4]（*Minimum Error Rate Training*，*MERT*）算法来获取最优的特征权重。该方法的主要思想是以翻译错误率为目标，在整个特征空间中，先固定其他维的权重，通过线性搜索方法调整每一维的特征权重。然而，由于训练语料的领域不尽相同，在特定开发集上训练得到的最优权重，在测试集上不一定是最优的，也就是存在过拟合问题。

为了解决*MERT*训练过拟合问题，研究者们提出了许多切实可行的方法。Song等人[5]利用集成学习*Bagging*的方法构造出*N*个不同的开发集，并且分别训练*N*个子系统，解码时利用子系统的结果投票产生最终译文。Duh和Kirchhoff[6]将*MERT*作为一个弱学习器（*weak learner*），然后使用*Boosting*方法对*N-best*译文重排序来得到最优译文。Xiao等人[7]同样基于*Boosting*方法，根据翻译结果的好坏对开发集语料赋予不同的权重，翻译结果较差的句子将得到更高的权重，使得下一轮迭代过程中翻译系统对句子进行更为充分的训练，以得到更优的系统，最后集成多个子系统得到一个翻译能力强的集成系统。王志洋等人[8]使用*MERT*过程产生的各轮中间结果，也就是特征权重，对验证集进行解码，并根据结果对各轮参数权重进行投票，从而获得更稳定和准确的参数。最近，黄书剑等人[9]和Clark[10]等人均采用非线性模型来对统计机器翻译进行建模，以避免传统线性模型的缺陷。

 与现有方法不同，本文不是构建翻译子系统，而是直接将集成学习思想融入到最小错误率训练过程中来解决训练过拟合问题。具体而言，本文方法通过选取*N-best*译文对应的特征子集来训练该子集对应的权重，然后使用每组特征子集的*BLEU*值得分来对权重进行加权平均，最终得到一组能够有效避免训练过拟合问题的权重。

 本文的组织结构如下：第2节介绍最小错误率训练过程；第3节将对我们提出的基于集成学习思想的最小错误率训练算法进行详细描述；第4节，我们通过实验来考察所提出方法的效果，第5节将对本文进行总结并对未来工作进行展望。

**2 最小错误率训练**

现有的大部分统计机器翻译系统以对数线性模型建模，最小错误率训练算法则是它的标准调参方法。一般地，对于一个汉语句子，经翻译系统翻译，会得到一个候选译文的集合，最优候选译文可用如下式子表示：

  (1)

其中，代表对数线性模型的特征函数，代表特征函数权重。假设可以用函数来表示机器生成译文对于参考译文的错误个数，并且错误对于多个句子是可以累加的：。那么，最小错误率训练方法的目标就是通过调节特征权重，使得开发集的总体错误率最低。在具体的实现过程中，该方法通常以*BLEU*值的负值作为目标函数，采用固定其它维权重，而对一维权重进行调整的方法来获取最终权重。下面将分别对*BLEU*评价指标和调参过程进行简单介绍。

**2.1 *BLEU*评价指标**

 *BLEU*[11](bilingual evaluation understudy)是一种机器翻译译文自动评价指标。其核心思想是“机器翻译的结果与专业人工的翻译越接近，则它的质量就越高”。*BLEU*采用*n-gram*匹配度来衡量一个机器翻译结果与专业人工翻译的相似度。*n-gram*的准确度定义如下：

  (2)

其中，*c*为候选译文集合中的句子，表示一个*n-gram*片段在参考译文中出现的最大匹配次数，表示候选句子中*n-gram*的总数。

 为了防止较短译文的得分过高，得分还要乘以一个长度罚分比*BP*（*brevity penalty*），*BP*及最后得分定义如下：

  (3)

  (4)

公式3中，表示候选译文单词数，表示参考译文中单词数与最接近的句子的单词数。公式4中，表示各*n-gram*匹配准确度的权重，取值为，为*n-gram*的匹配准确度。*BLEU*评价指标已被证明与人工评测有较高的相关性。

**2.2最小错误率训练方法**

最小错误率训练方法的流程包含下列几个步骤：

1. 首先，初始化特征权重并进行解码，产生*N-best*特征文件；
2. 然后，通过最小错误率训练得到新的特征权重，并用这组特征权重进行解码得到新的*N-best*特征文件；
3. 将新生成的*N-best*译文特征和之前的*N-best*译文特征文件合并，再在扩展的*N-best*译文特征上重新调参；
4. 如此循环，直至没有新的*N-best*译文产生。

在进行参数训练时，*MERT*所进行的操作是先调节某一维的权重，而其它的特征权重将被看作常数。因此，对于指定的源语言句子，其候选译文的得分可以表示成：

  (5)

其中，是在解码过程获得的特征值，是当前要调节的特征权重。那么，将固定为常数时，候选译文*t*的得分就是关于待调整权重的线性函数，如图1中的曲线*a*。因此对于候选译文集中的每个候选译文，都对应存在同样的一个线性函数，如图1（假设中有，，三个候选译文）所示。

 由图1可知，随着的改变，最优译文选择只会在临界点和处发生变化，而在任意两个相邻的临界点之间，最优译文不发生变化。比如在区间上，译文的得分最高，为这段区间上的最优译文。那么针对一个源语言句子和对应解码得到的若干候选译文，利用公式5便可以计算出各个临界点。对于整个开发集而言，将所有源语言句子所对应的临界点合并后排序，那么在相应得到的相邻两个临界点之间，值的变化不会导致开发

**图 1 最优译文*e*关于特征权重变化曲线**

Figure 1 curves of optimal translation *e* fluctuating due to feature weight 

集中任何一个句子的最优译文选择发生变化。最后，计算每个区间内所有译文整体的*BLEU*值，就可以找到这一维变化上最优的权重。

 最小错误率训练是一个有效的调参方法，但是在训练过程中容易出现训练过拟合现象。针对这个问题，本文提出基于集成学习的最小错误率训练算法。下面章节将介绍本文所提出的方法。

**3 基于集成学习的最小错误率训练算法**

本章将首先介绍集成学习方法的基本思想，然后描述如何将集成学习融入到最小错误率训练方法中。

**3.1 集成学习**

 集成学习（*Ensemble Learning*）是一种通用的机器学习框架。它使用多个学习机来解决同一问题,由于该框架在统计上、计算上和表示上都具有显著的优势，因此它能够显著提高学习系统的泛化能力[12]。早在1997年，国际机器学习届的权威T. G. Dietterich就将集成学习列为了机器学习四大研究方向之首[13]。目前，集成学习已经广泛应用于Web信息过滤、生物特征识别、计算机辅助医疗诊断等众多领域，可以说能用到机器学习的地方，就能用到集成学习。

 集成学习使用多个学习机来解决同一问题，一般地，一个集成学习机的构建分为两步：基学习机的构建和基学习机的集成。构建基学习机的方法又大致可以分为两类：一类是将不同类型算法应用于同一

|  |
| --- |
| **Algorithm 1 *MERT* Based on Ensemble Learning** |
| **Input:** a training set {(,),,(,)} where is the *i*-th source sentence and  is the set of reference translations for , the complete set of features , the set of initialized *feature weights* **Output:** a new set of *feature weights* **Define:** *M* is the number of features, and *MERT* procedure iterates *T* times |
| 1: **for** each iteration **do:** 2: **for** **do:** 3: **extract** subset of features 4: **extract** subset of feature weights 5: **train** to update the feature weights set  and get its bleu score 6: **set** , where  is set to **0.**7: **end for**8: **calculate** the Euclidean distance between any two weight vectors.9: **filter** the weight vectors that have large distance to others.10: , where  contains the rest of weight vectors of small deviation after filtering step11: **normalize** the weight vector 12: **end for** |

图2 基于集成学习的最小错误率训练

Figure 2 *MERT* based on Ensemble Learning

训练集，得到的基学习机通常被称为是异质类型的（*heterogeneous*）；另一类是将同一算法应用于不同的训练集，这种方法得到的基学习机被称为是同质类型的（*homogeneous*）。对于生成同质类型的基学习机，以*Boosting*[14]和*Bagging*[15]方法影响最大。而合并集成系统的方法也有很多，例如简单投票、贝叶斯投票、基于D-S证据理论的整合方式等。

**3.2 基于集成学习的最小错误率训练**

 本文的集成方法属于同质集成学习方法。具体而言，本文首先采用对数线性模型的*M*个特征函数作为特征全集，通过每次减少一个特征的方法，构造出*M*个特征子集：,其中代表不包含第*i*个特征的一组特征子集。

 在特征子集构造完成之后，就是本文所提出的基于集成学习的最小错误率训练过程：将整个*MERT*方法看作一个函数，其输入是开发集（一个源语言文件和若干参考译文文件）以及特征配置文件（包含特征集合和初始化的特征权重），输出是这一轮训练中最优结果对应的*BLEU*值和特征权重。那么在每一轮迭代过程中，分别用*M*组特征子集进行最小错误率训练，便可以得到*M*组最优权重及其对应在开发集上的*BLEU*值。然后，用每组特征权重对应的*BLEU*值作为权重，对所有的特征权重进行加权集成（缺省的一维特征权重为0），以得到最终的特征权重。这组权重将作为初始值在下一轮迭代过程中继续执行上述参数调节过程。图2算法1描述了这一过程。首先，抽取出减少了一个特征的特征子集以及相应的初始权重（算法1第3-4行），然后用*MERT*方法训练这组特征子集对应的特征权重并将缺少的特征维度的权重设为0（算法1第5-6行），在得到M组新的权重向量之后，计算其两两之间的欧氏距离（算法1第8行），剔除与其它组向量差异较大的权重向量，而只对其中K组偏差较小的权重进行加权集成（算法1第9-10行，具体约束方法将在3.4节介绍），最后对权

图 3 迭代过程中训练得到的九组权重示例

Figure 3 nine example weight vectors generated by training

表 1 各组权重向量分别与其余所有权重向量间的欧氏距离

Table 1 Euclidean distance between each weight vector and each others

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
|  | 0 | 0.117 768 | 0.150 093 | 0.553 534 | 0.572 255 | 0.344 292 | 0.155 867 | 0.371 206 | 0.162 071 |
|  | 0.117 768 | 0 | 0.081 915 | 0.519 196 | 0.548 091 | 0.308 478 | 0.059 268 | 0.313 156 | 0.201 628 |
|  | 0.150 093 | 0.081 915 | 0 | 0.499 814 | 0.531 395 | 0.349 869 | 0.088 817 | 0.354 226 | 0.218 416 |
|  | 0.553 534 | 0.519 196 | 0.499 814 | 0 | 0.125 287 | 0.565 943 | 0.535 824 | 0.559 005 | 0.510 770 |
|  | 0.572 255 | 0.548 091 | 0.531 395 | 0.125 287 | 0 | 0.583 628 | 0.560 945 | 0.575 403 | 0.516093 |
|  | 0.344 292 | 0.308 478 | 0.349 869 | 0.565 943 | 0.583 628 | 0 | 0.300 604 | 0.121 830 | 0.337 047 |
|  | 0.155 867 | 0.059 268 | 0.088 817 | 0.535 824 | 0.560 945 | 0.300 604 | 0 | 0.295 906 | 0.227 981 |
|  | 0.371 206 | 0.313 156 | 0.354 226 | 0.556 005 | 0.575 403 | 0.121 830 | 0.295 906 | 0 | 0.350 657 |
|  | 0.162 071 | 0.201 628 | 0.218 416 | 0.510 770 | 0.516 093 | 0.337 047 | 0.227 981 | 0.350 657 | 0 |

此处第5组和第11组特征分别为语言模型特征和词惩罚特征，根据3.3节所述，不予以考虑，故此处不列出。计算所有距离的平均值即距离阈值σ为0.352，经过比较可知由第4、6组权重向量计算出的欧氏距离中，分别都有7个距离超过了阈值σ，按照约束规则(2)，该两组权重向量将不参与集成过程。

重向量进行归一化（算法1第11行），并将得到的特征权重用于下一轮迭代过程。

值得注意的是，在统计机器翻译的对数线性模型当中，通常包括两个很重要的特征：语言模型特征和词惩罚特征。语言模型特征对任何一个统计机器翻译系统都是必不可少的，它反映出译文的流利程度；词惩罚特征主要用于平衡语言模型的负面影响，因为语言模型会倾向于用长度短的译文，词惩罚能够有效调整译文的长度。这两个特征对系统翻译质量的好坏有着重要影响，缺少这两个特征的系统的翻译效果很差。为了使最后集成的系统得到最优效果，在集成的过程中，本文不对这两组特征权重进行加权，即强制要求所有的特征子集都要包含上述两个特征。

**3.3 启发式约束**

 在对特征权重集成的过程中存在一个问题：在同一轮迭代过程中产生的两组特征权重向量，虽然其对开发集的翻译效果相近（例如*BLEU*值几乎相等），但是特征权重在某一维（或某几维）上有较大的偏差，最后加权平均得到的特征权重会偏离最优权重较远。这一问题，在训练过程的开始几轮中更为明显，因此，不合适直接应用集成学习方法。

针对这一问题，本文采用如下两条约束规则：

1. 为了避免由于训练过程的开始几轮，特征权重偏离较大而产生的影响，本文在第轮迭代过程之后再应用基于集成学习的最小错误率方法。也就是，在前轮迭代过程中，使用初始的*M*个特征作为特征全集进行*MERT*训练，之后再运用本文所提出的方法（实验中，经验性地将设置为8）。
2. 为了集成得到更好的特征权重，本文通过计算特征权重之间的欧氏距离来筛选待集成的权重。在一轮迭代中，对于一组特征权重，计算它与其它各组权重之间的欧氏距离，若得到的这组距离中有半数以上超过设定的距离阈值，那么认为这组特征权重偏差较大，不适合参与集成。实验中，本文设置距离阈值为所有特征权重之间的欧式距离的平均值。

 关于这一筛选过程，举例如下，假设实验在某一轮迭代过程中得到了如图3所示的权重向量。可见，在第3个特征维度上，各特征权重值偏差较大，如果不对权重向量进行选择，那么全部集成后的权重会更加偏离最优权重。经过计算，可以得到如表1所示各权重向量间的欧式距离，并可以求得所有距离的平均值即距离阈值*σ*为0.352，按照第(2)条约束规则，本文不对其中的4、6组权重向量进行集成。

**4 实验**

**4.1 实验设置**

 本文通过*NIST*新闻领域评测和*IWSLT*口语领域评测中英机器翻译任务来考察本文方法的有效性，实验设置如下：

1) 训练语料：对于*NIST*任务，本文使用*LDC*[[2]](#footnote-3)双语平行语料。该语料库共含有1百万个平行句对；对于*IWSLT*任务，本文使用*BTEC*[[3]](#footnote-4)双语平行语料。该语料库共含有4万个句子。其中，对于中文句子，本文使用*ICTCLAS*[[4]](#footnote-5)进行分词；而对于英文句子，本文则使用由Koehn等人开发的欧洲语分词脚本进行分词处理[[5]](#footnote-6)。在做完预处理之后，本文利用开源的词语对齐工具*GIZA++*[[6]](#footnote-7)对分词后的语料进行词语对齐。在此过程中，本文所使用的对齐合并方式是“*grow-diag-final-and*”[16]。

2) 语言模型：对于*NIST*任务，本文使用同样来自*LDC*的*Giga*新华语料训练一个4元的语言模型；对于*IWSLT*任务，本文则使用*BTEC*的英文部分训练一个4元的语言模型。具体实现中，本文使用由Koehn等人开发的欧洲语分词脚本进行分词处理，使用的训练工具是*SRILM*[[7]](#footnote-8)。

3) 开发集和测试集：对于*NIST*任务，本文使用*NIST05*作为开发集，*NIST06和*

表2 *NIST*翻译任务中采用不同训练方法时系统的*BLEU*值（\*和\*\*分别表示*P*<0.05和*P*<0.01的

显著性差异）

Table 2. *BLEU* scores of translation system using different training methods in *NIST* translation task（\* and \*\*: significantly better than baseline at p<0.05 and p<0.01, respectively.）

|  |  |  |  |
| --- | --- | --- | --- |
| 解码器 | 训练方法 | *NIST06* | *NIST08* |
| 短语系统 | Baseline | 28.12 | 21.32 |
| Ensemble Learning without Heuristic Method | 28.20 | 20.90 |
| Ensemble Learning with Heuristic Method | 28.79\*\* | 22.30\*\* |
| 层次短语系统 | Baseline | 30.25 | 21.25 |
| Ensemble Learning without Heuristic Method | 30.54 | 21.63\* |
| Ensemble Learning with Heuristic Method | 30.60\* | 21.77\* |

表3 *IWSLT*翻译任务中采用不同训练方法时系统的*BLEU*值

Table 3. *BLEU* scores of translation system using different training methods in *IWSLT* translation task

|  |  |  |  |
| --- | --- | --- | --- |
| 解码器 | 训练方法 | *IWSLT05* | *IWSLT07* |
| 短语系统 | Baseline | 55.70 | 32.32 |
| Ensemble Learning without Heuristic Method | 55.12 | 32.45 |
| Ensemble Learning with Heuristic Method | 56.66\*\* | 33.51\*\* |
| 层次短语系统 | Baseline | 57.32 | 33.86 |
| Ensemble Learning without Heuristic Method | 57.45 | 34.57\*\* |
| Ensemble Learning with Heuristic Method | 57.92\*\* | 34.68\*\* |

*NIST08*作为测试集；对于*IWSLT*任务，本文使用*IWSLT04*作为开发集，*IWSLT05和IWSLT07*作为测试集。

4) 解码器：我们在短语系统*MOSES*[[8]](#footnote-9)*和*层次短语系统上验证本文方法的有效性，下一节将介绍所采用的解码器。

5) 评测指标：采用大小写不敏感的*BLEU*评测指标（计算方法已在2.1节介绍），使用的评测工具是*mteva-v11b.pl*[[9]](#footnote-10)。并对译文进行差异性检测[17]。

**4.2 *解码器***

 本文中主要使用短语解码器[18]和层次解码器[19]来验证本文方法的有效性。

本文使用的短语解码器是MOSES。*MOSES*是著名的开源统计机器翻译系统，采用对数线性模型。在实验中，该系统使用下列特征：(1) 双语短语的源语言-目标语言的规则翻译概率；(2) 双语短语的目标语言-源语言的规则翻译概率；(3) 双语短语的源语言-目标语言的规则词汇化权重；(4) 双语短语的目标语言-源语言的规则词汇化权重；(5) 短语惩罚分数；(6) 语言模型权重；(7)词惩罚分数；(8) 在***monotone***方向上的调序概率；(9) 在***swap***方向上的调序概率；(10) 在***discontinous***方向上的调序概率；(11) 扭曲距离。

本文作者重现了层次短语解码器[14]。该系统同样采用对数线性模型，包含下列特征：(1) 双语规则的源语言-目标语言的规则翻译概率；(2) 双语规则的目标语言-源语言的规则翻译概率；(3) 双语规则的源语言-目标语言的规则词汇化权重；(4) 双语规则的目标语言-源语言的规则词汇化权重；(5) 规则惩罚分数；(6) 语言模型权重；(7) 词惩罚分数；(8) 粘合规则惩罚分数。

**4.3 实验结果**

 本文进行了三组实验：第一组是使用最小错误率训练方法，也是本文的基线系统；第二组是应用了集成学习思想的系统，但没有进行阈值过滤；第三组是应用了集成学习思想并进行阈值过滤的系统。特别地，在第三组实验中，本文经验性地设置权重的距离过滤阈值为所有距离的平均值。

表2给出了*NIST*翻译任务的实验结果。使用传统最小错误率训练方法，*MOSES*在*NIST06*，*NIST08*的*BLEU*值分别为28.18, 21.32。对于未经过阈值过滤得到的集成权重，其在*NIST06*、*NIST08*语料上的*BLEU*值分别为28.20，20.90，并没有优于基线系统。在第三组实验中，本文对参与集成的特征权重根据空间距离进行了过滤，然后再进行集成。实验结果表明，使用该训练方法的系统在两个测试集合上*BLEU*值具有一定程度的提高。具体而言，采用新训练方法，*NIST06*上系统*BLEU*值为28.79，*NIST08*上系统*BLEU*值为22.30，相比基线系统，采用新训练方法的系统分别提高了**0.61**和**0.98**。同样，层次短语系统在采用本文方法之后，译文质量也获得了提高，相比基线系统，采用基于阈值过滤的集成学习训练方法的系统*BLEU*值分别提高了**0.35**和**0.52**。

表3给出了*IWSLT*翻译任务的实验结果。使用传统最小错误率训练方法，*MOSES*在*IWSLT05*，*IWSLT07*的*BLEU*值分别为55.70和32.32。与*NIST*翻译任务的实验结果相似，没有使用空间距离进行阈值过滤而集成得到的权重，并没有显著提高系统的翻译性能，甚至在*IWSLT06*还出现了显著的性能下降。而采用空间距离进行阈值过滤而集成得到的权重，系统性能均有显著提高。在两个测试集上，采用新训练方法的系统*BLEU*值分别为56.66和33.51，该结果相比基线系统分别提高了**0.96**和**1.19**。与前面实验结果相同，层次短语系统在采用本文方法之后，译文质量也获得了提高，相比基线系统，采用基于阈值过滤的集成学习训练方法的系统*BLEU*值分别提高了**0.6**和0.82。

上述结果证明了本文所提方法对于解决传统最小错误率训练的过拟合问题具有较好的效果。特别地，在*NIST*翻译任务中，新方法在*NIST08*上的实验效果相比其他集合的提高更为明显，因此，可认为本文方法对于领域不相同的文本翻译具有更好的效果。

**5 结论及展望**

本文将集成学习的思想直接融入到统计机器翻译的最小错误率训练过程中。在*NIST*和*IWSLT*上的实验结果表明，本文方法具有较好的效果。

 但是，本文对特征子集的选取非常简单，在接下来的工作中我们将进一步探讨特征子集的选取对集成系统翻译效果的影响，考虑更加复杂的选取策略以得到更优的集成系统。同时，集成学习过程需要训练多个子系统，训练时间及所需的存储空间也迅速增加，如何平衡训练的时空存储和翻译质量的问题也是需要进一步思考的问题。

**参考文献（References）：**

[1] KOEHN P. **Statistical Machine Translation** [M]. Beijing: Publishing House of Electronics Industry, 2012

[2] 刘群. 统计机器翻译综述 [J]. 中文信息学报，2003

[3] OCH F J, Ney H. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation [C]. **Proc of ACL 2002.** Philadelphia: Association for Computational Linguistics, 2002: 295-302

[4] OCH F J. Minimum Error Rate Training in Statistical Machine Translation [C] **Proc of ACL 2003.** Sapporo, Japan: Association for Computational Linguistics, 2003: 160-167.

[5] SONG L, Mi H, Lü Y. Bagging-based System Combination for Domain Adaptation [C]. **In Proc of the** **MT Summit XIII**, 2011

[6] DUH K, KIRCHHOFF K. Beyond log-linear models: boosted minimum error rate training for n-best re-ranking [C]. **Proc of ACL 2008**:HLT, Short papers

[7] XIAO Tong, ZHU Jing-bo, ZHU Mu-hua, et al. Boosting-based System Combination for Machine Translation [C]. **Proc of ACL 2010.** Uppsala: Association for Computational Linguistics, 2010: 739-748

[8] 王志洋，姜文斌. 基于投票平均的最小错误率训练算法 [C]. 第五届全国机器翻译研讨会. 中国，南京，2009

[9] Shujian Huang, Huadong Chen, Xinyu Dai and Jiajun Chen. Non-linear Learning for Statistical Machine Translation [C]. To appear in **Proc of ACL** 2015. Association for Computational Linguistics, 2015.

[10] Jonathan H. Clark, Chris Dyer and Alon Lavie. Locally Non-linear Learning for Statistical Machine Translation via Discretization and Structured Regularization [J]. Translation **of ACL** 2014. Association for Computational Linguistics, 2014(2): 393-404.

[11] PAPINENI L, ROUKOS S, WARD T, et al. BLEU: a Method for Automatic evaluation of Machine Translation Evaluation [C]. **Proc of ACL 2002.** Philadelphia: Association for Computational Linguistics, 2002: 311-318.

[12] 梁英毅. 集成学习综述 [EB/OL]. 2006

[13] DIETTERICH T G. Machine learning research：Four current directions. **AI Magazine**,1997,18(4)：97-136

[14] FREUND Y, SCHAPIRE RE. A decision-theoretic generalization of on-line learning and an application to boosting [J]. **Journal of Computer and System Sciences**, 1997, 55(1)：119-139

[15] BRREIMAN L. Bagging predictors. **Machine Learning** [M], 1996, 24(2)：123-140

[16] OCH F J, NEY H. A systematic comparison of various statistical alignment models [J]. **Computational Linguistics**, 2003: 19-51

[17] KOEHN P. Statistical Significance Tests for Machine Translation Evaluation[C]. **Proc of EMNLP 2004.** Barcelona, Spain: Association for Computational Linguistics, 2004: 388-39

[18] KOEHN P, OCH F J, MARCU D. Statistical Phrase-Based Translation [C]. **Proc of** **HLT-NAACL 2003**. Edmonton, Canada: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2003: 48-54

[19] David Chiang. Hierarchical Phrase-based Translation [J]. Computational Linguistics, 2007: 201-228

1. **收稿日期：**

**作者简介：**陈 昉 （1980-），男，主要从事自然语言处理研究。Email:softtnet@xmu.edu.cn赵程绮（1993-），男，主要从事自然语言处理研究。Email：zcq@stu.xmu.edu.cn

王志豪（1993-），男，主要从事自然语言处理研究。Email：wzh@stu.xmu.edu.cn

 李江梦（1993-），男，主要从事自然语言处理研究。Email：jml@stu.xmu.edu.cn [↑](#footnote-ref-2)
2. *LDC*语料：http://www.ldc.upenn.edu [↑](#footnote-ref-3)
3. *BTEC*语料：http://iwslt2010.fbk.eu/ [↑](#footnote-ref-4)
4. *ICTCLAS*工具：http://ictclas.org/ [↑](#footnote-ref-5)
5. 分词脚本：http://www.statmt.org/ [↑](#footnote-ref-6)
6. *GIZA++*工具：http://code.google.com/p/giza-pp/ [↑](#footnote-ref-7)
7. *SRILM*工具：http://www.speech.sri.com/projects/srilm/download.html [↑](#footnote-ref-8)
8. *MOSES*：http://www.statmt.org/moses/ [↑](#footnote-ref-9)
9. *BLEU*：http://www.nist.gov/speech/tests/mt/resources/scoring.hml [↑](#footnote-ref-10)